巨量資料與統計分析

政治大學統計系余清祥

2024年10月15日

第六週：結構資料分析

http://csyue.nccu.edu.tw

常見的分析方法

關於結構資料的分析

- 結構資料的變數及其欄位已有定義，確定資料品質、偵錯之後，其分析方法較為具體：

→若給定目標變數，可從所有可能的排列組合找出最佳者，類似迴歸分析中的Mallow Cp。

- 資料量使分析更形複雜，可先考慮EDA（基本統計量、圖表）找出大略趨勢，或抽樣（交叉驗證）探索資訊，再套用CDA進一步分析。

註：參考「R and Data Mining: Examples and Case Studies」。

常見的結構資料分析方法

- Techniques (參考*The Elements of Statistical Learning*)

→分類(Classification)與群聚分析(Cluster Analysis)

→羅吉士迴歸(Logistic Regression)

→分類樹(Classification and Regression Tree；CART)

→類神經網絡(Neural Networks；NN)

→支持向量機(Support Vector Machine；SVM)

→無母數迴歸(Nonparametric Regression)

→時間序列(Times Series)

→密度估計(Density Estimation)

羅吉士迴歸(Logistic Regression)

- 羅吉士迴歸用於處理二元格式(記為0和1)的目標變數，操作與詮釋與一般迴歸相當接近。
- 羅吉士函數(又稱S型或反曲函數) $f(x)=\frac{1}{1+e^{-x}}=\frac{e^{x}}{1+e^{x}}$

  將一般迴歸式 $y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$

  右側以羅吉士函數帶入，或是

$$\log(\frac{p}{1-p})=\beta_0+\beta_1X_1+\cdots+\beta_kX_k;\quad 或是\ y=\frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}$$

  許多人以**0.5**當成分界的門檻值(Threshold)

- 參數 β_i 要透過勝算比(Odds Ratio)概念呈現，亦即亦即 每增加一個單位X_i對整體Y增加/減少的機率

→每天多抽一根香菸，罹患肺癌及頭頸癌機會增加60%

分類樹(CART)

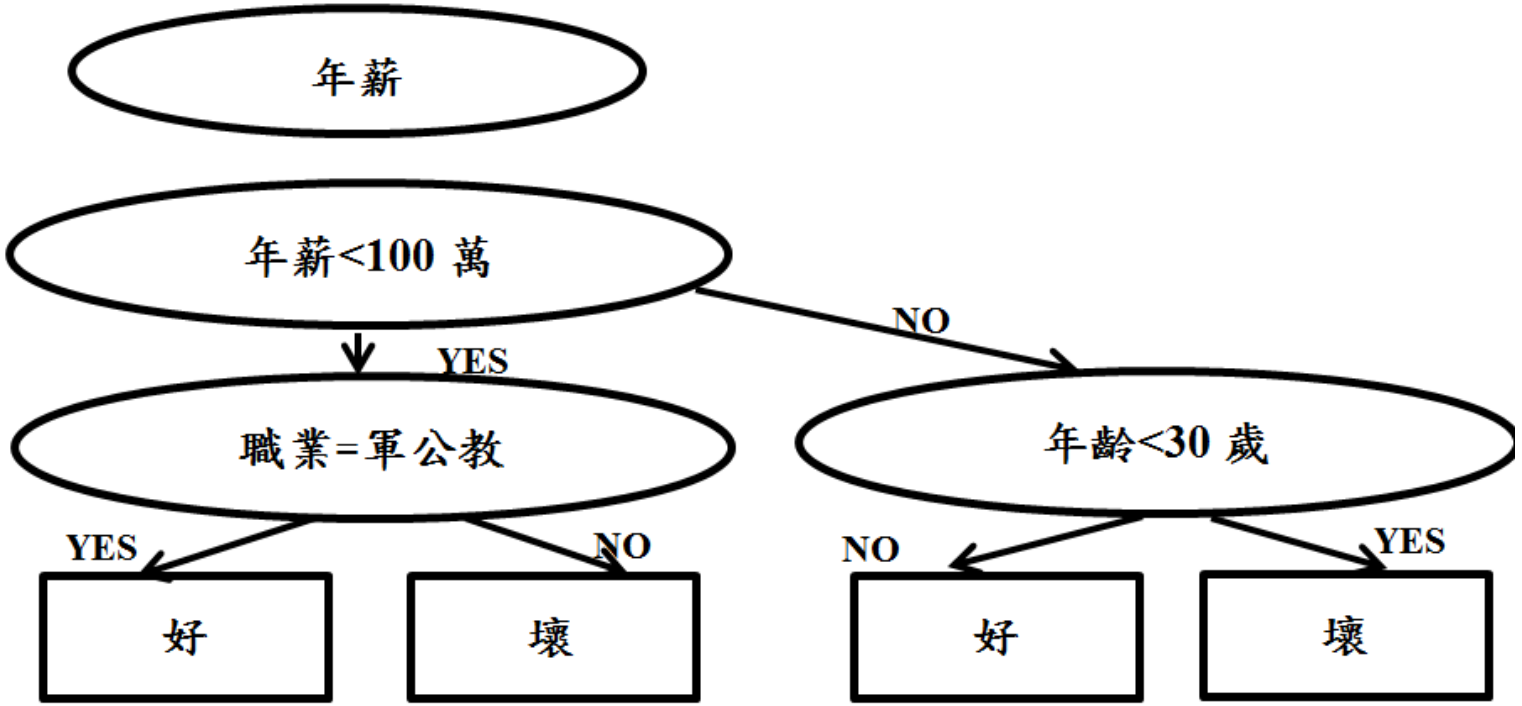
- **分類樹**又稱決策樹(Decision Tree)，是運籌學(作業研究)常用的決策分析工具，應用於信用評等、醫學檢查、醫療處方等。
- 特色為**視覺圖像化**方式，將所有可能分析結果以**樹狀圖**呈現，讓人很容易一目了然，藉此協助研究者快速決定最可行方法。

→以發生機率較高、支出費用較少等屬性作為判斷目標

- 分類樹流程圖有一個反應變數和一個以上的解釋變數，**每個分枝節點均為一個二元試驗**，以分支規則決定樣本送到下層節點方式。
- 缺點為忽略變數間關聯性、如何處理遺漏值？

分類樹(CART)-銀行評估信用評比

- 年薪是最重要的評估變數，決策流程由此開始。
- 年薪<100萬是**根部節點**(Root Node)作為資料分野
- 職業、年齡是**中間節點**(Non-Lead Node)是條件判斷
- 方框是信用評比為**葉節點**(Leaf Node)，完成分類標記



類神經網絡(NN)

- NN又稱人工神經網路，仿照生物體神經元(Neuron)的組成與傳送結構。當神經元接受外部刺激或其他神經元傳遞訊息（接受刺激），經過簡單處理（計算）後將執行結果傳遞外界或其他神經元（傳遞反應）。

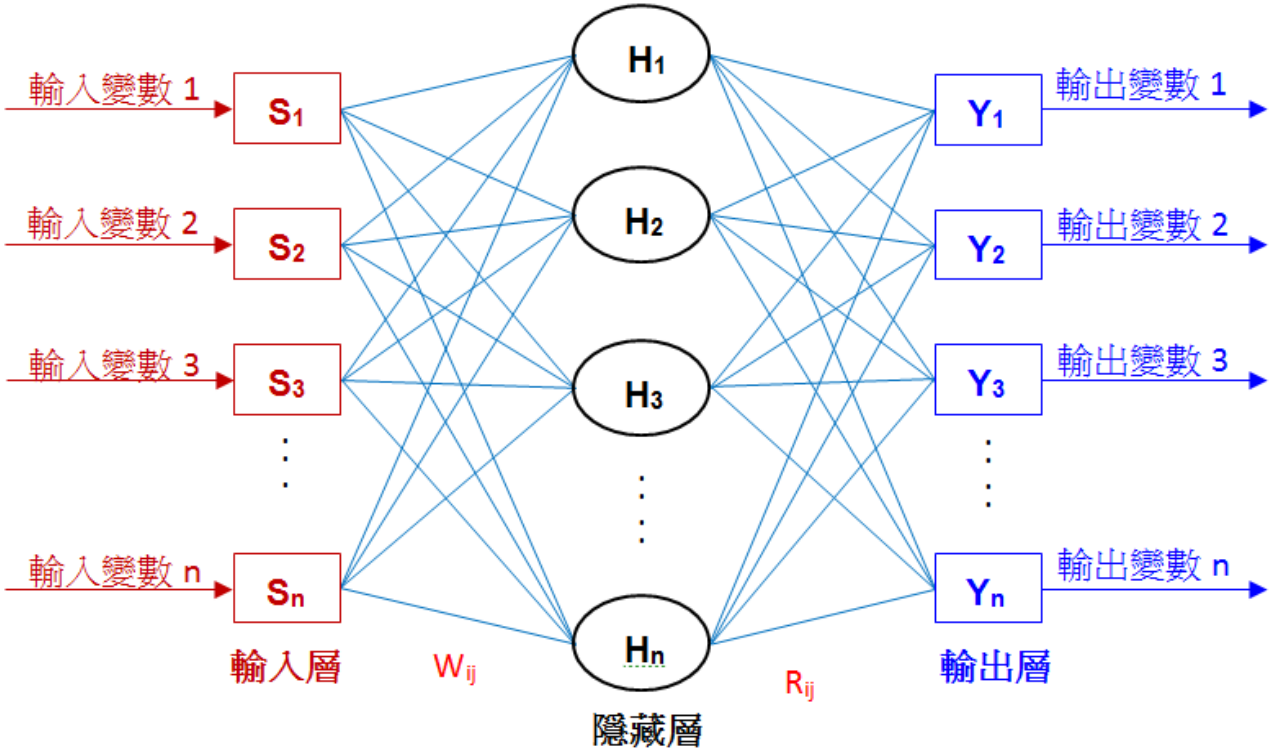
→生物體機能受損時，神經元會重新學習/計算。類神經網絡也相同，會不斷訓練/測試，因此計算時間較久

- 應用在判斷信用卡盜刷、股票指數預測、基因演算法智慧型辨識系統（如臉部辨識等）、汽車控制、家電與機器控制等。AlphaGo也是知名應用案例。

類神經網路(NN)的三個分層

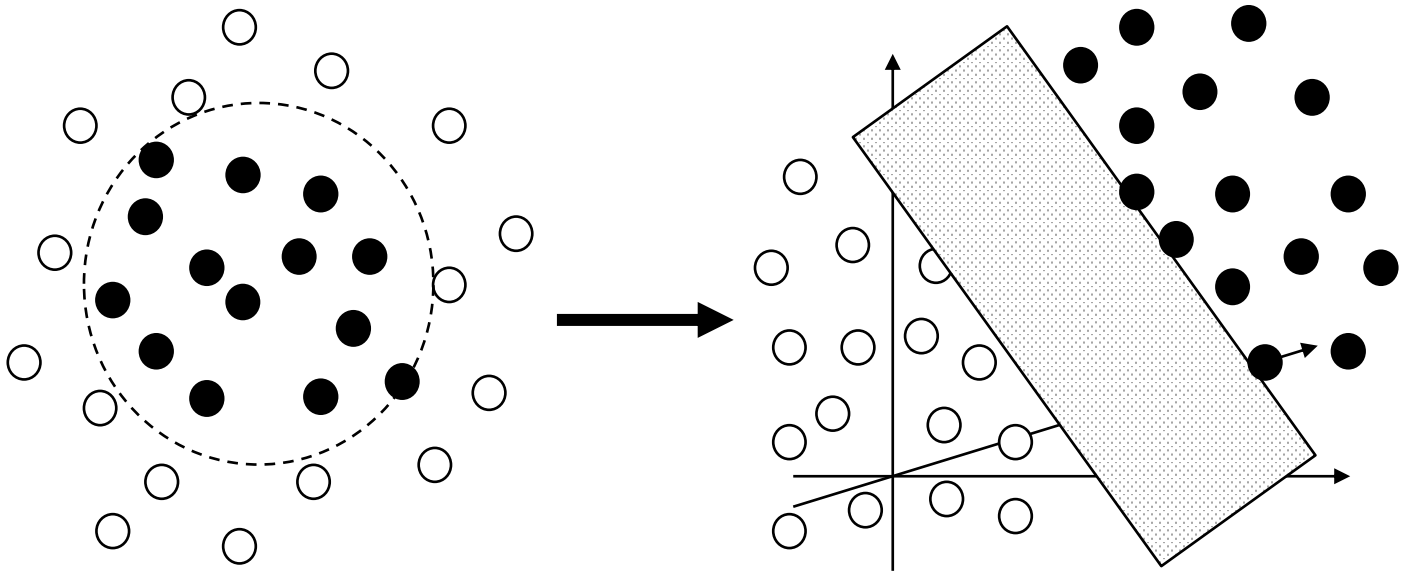
- 基本架構：**輸入層**(Input Layer)、**隱藏層**(Hidden Layer)、**輸出層**(Output Layer)三層。

→中間為隱藏層，可能有一至多個隱藏層，每層均有數個神經元互相連接，透過調整神經元權重，使模型趨於**收斂**。



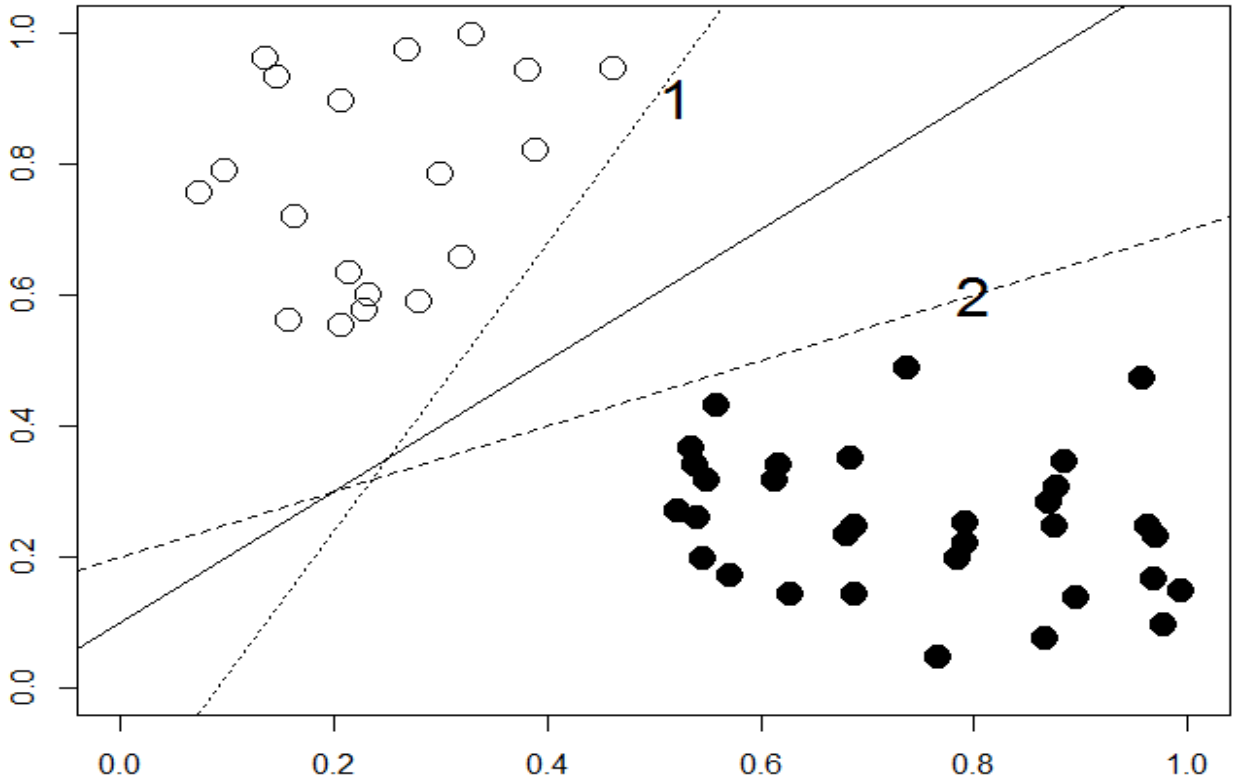
支持向量機(SVM)

- SVM概念和資料縮減(Data Reduction)相反，將原始資料映射到高維度空間後，再用資料用簡單分類函數型態，像是一直線或一超平面(Hyperplane)分離觀察值。



支持向量機(SVM)的最大邊界

- SVM以達到**最大距離邊界**(Margin)為目標，也就是分類結果中兩類觀察值到邊界的最小距離盡量越大越好，才可讓分類錯誤率降到最低。



案例一、癌症死亡率

案例一、癌症死亡率

- 臺灣自1995年實施**全民健康保險**(National Health Insurance; NHI)，現在約有99%人口納保，涵蓋臺灣全體人民，可不須由普查、調查蒐集健康議題資料。
- 設計出發點為**社會保險**、**風險分攤與共享**機制。

→**保費減免**(全額或部分)：中低收入戶、身心障礙者等

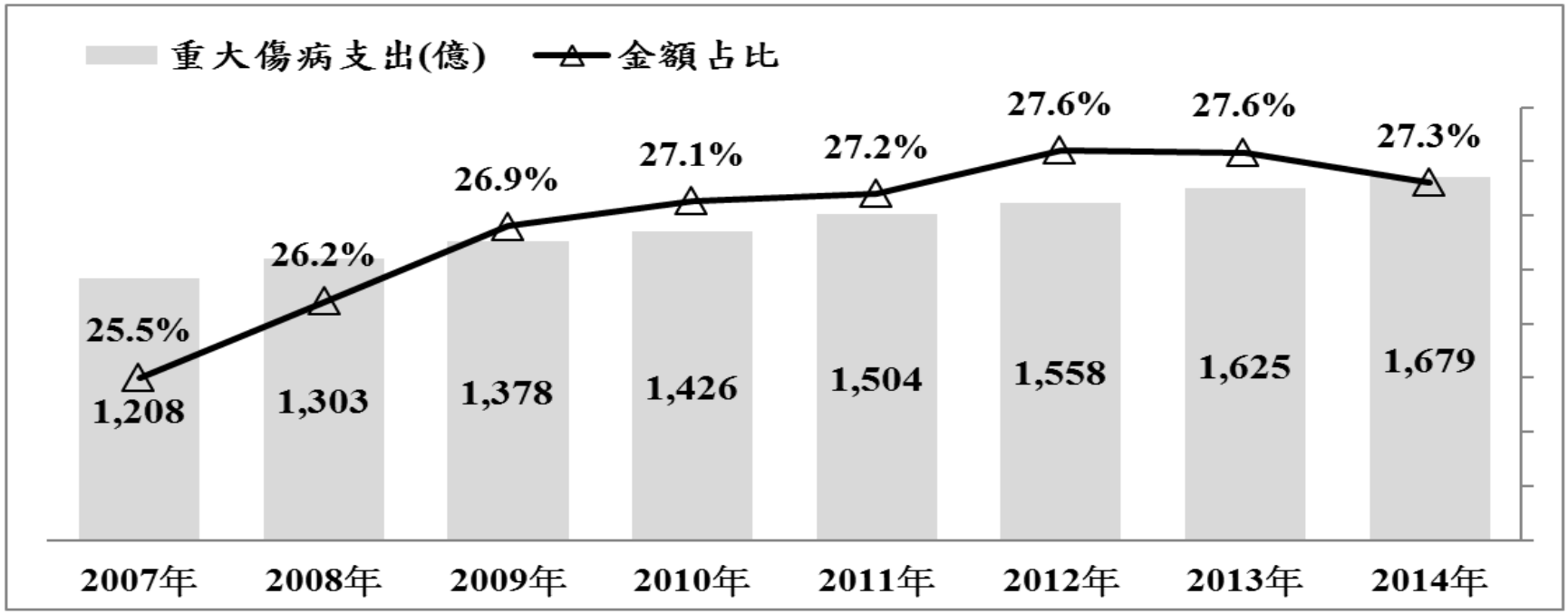
→**免部分負擔**(Waiver of copayment)：按照身份別(如：榮民榮眷、低收入戶、孕婦、離島居民等)或疾病別(如重大傷病患者)享有優惠。

- 可到主管機關**衛福部中央健康保險署**網站蒐集資料。

重大傷病占健保主要支出

- 2014年重大傷病患者約佔人口的4%，但其醫療費用卻佔所有健保支出的27.3%。
- 重大傷病發生率、盛行率與年齡成正比。

→臺灣人口老化嚴重，重大傷病支出將持續擴大！



癌症為重大傷病人數之首

排序	疾病別	領證數	占有率
1	需積極或長期治療之癌症	442,871	46.0%
2	慢性精神病	202,653	21.1%
3	需終身治療之全身性自體免疫症候群	102,284	10.6%
4	慢性腎衰竭（尿毒症），必須接受定期透析治療者	79,576	8.3%
5	心、肺、胃腸、腎臟、神經、骨骼系統等之先天性畸型及染色體異常	35,641	3.7%
6	先天性新陳代謝異常疾病	14,238	1.5%
7	小兒麻痺、腦性麻痺所引起之神經、肌肉、骨骼、肺臟等之併發症者（其身心殘障等級在中度以上）	12,916	1.3%
8	接受腎臟、心臟、肺臟、肝臟、骨髓及胰臟移植後之追蹤治療	12,893	1.3%
9	因呼吸衰竭需長期使用呼吸器者	11,679	1.2%
10	罕見疾病	9,211	1.0%
其他		38,672	4.0%
合計		962,634	100%

資料來源：衛生福利部中央健康保險署2016年6月領有全民健保重大傷病的前10大疾病

健保資料庫的範例：

- 本案例串聯以下三檔資料。下表可發現「HV和HV_CD檔人數差異逐年擴大」，與「重大傷病需持續就醫享有免部分負擔補助」互相矛盾，**可能原因**為何？**怎麼檢查**這個數字？

Year	承保資料檔 (ID)		重大傷病證明明細檔 (HV)		重大傷病門診處方及治療明細檔 (HV_CD)	
	人數	人次	人數	人次	人數	人次
2007	26,421,498	26,619,007	1,103,431	1,453,483	649,106	17,946,211
2008	26,780,636	26,970,225	1,164,465	1,529,866	678,544	19,173,919
2009	27,067,952	27,223,008	1,276,315	1,733,251	712,828	20,357,173
2010	27,369,795	27,509,909	1,350,786	1,863,254	746,746	21,619,442
2011	27,699,442	27,841,406	1,401,449	1,933,455	779,179	22,861,178

由EDA確認資料品質與探索特性

- 接續施行EDA，計算重大傷病領證人數與人次，發現以下三個潛在訊息：

→HV檔的人次與人數兩者比值1.35，即平均一人有1.35張，推測有些病患領有兩張以上重大傷病卡（或證明、註記）。

→HV_CD檔的人次與人數比值為28.55，也從資料可看出平均每個重大傷病病患每年就診次數為27~30次。

→HV與HV_CD檔的人數並不相符，可能因病患已死亡或康復，但未註銷在HV檔的紀錄。（註：也有可能是病患領有多張重大傷病卡，或是重大傷病卡在住院時取得，並沒有經過門診的程序。）

- 如何確定病患已經死亡？怎麼從資料庫中判讀？

如何分析健保資料庫？

- 龐雜健保資料庫使得資料分析更為困難。

→需要仰賴資料庫軟體(e.g., SQL)及資料科學家(Data Scientists, e.g., IT experts)；

→資料清理通常耗費大量時間及人力，因為資料格式不統一（來自不同醫療院所）。

- 資料品質？

→因為大數據無法保證資料品質，需要根據問題目標重新挑選變數、整併資料。

使用健保資料庫計算癌症死亡率

- 大型資料庫中符合研究主題需求者，通常不只一種可能，研究者根據經驗及相關知識，選擇較為可行的子資料庫，設計及規劃進行步驟、分析方法。
- 以癌症死亡率為例，因為無法依賴既有資料欄位（資料品質），例如：健保資料庫中的死亡無法由「死亡註記」確定，必須仰賴其他方法判斷癌症患者是否死亡。

註：也可透過衛福部串連健保資料、死因檔。

以癌症死亡人數檢測資料品質

死亡註記	[空白]	0	1	2	3	4
人數	675,923	288,325	35,353	136,591	10,184	15,820
死亡註記	5	6	7	8	9	Y
人數	9,382	9,015	9,641	6,892	11,828	7,517

註：上述為2005年健保資料庫癌症死亡註記，太少！

與官方死於癌症人數（約四萬人）差異太大。

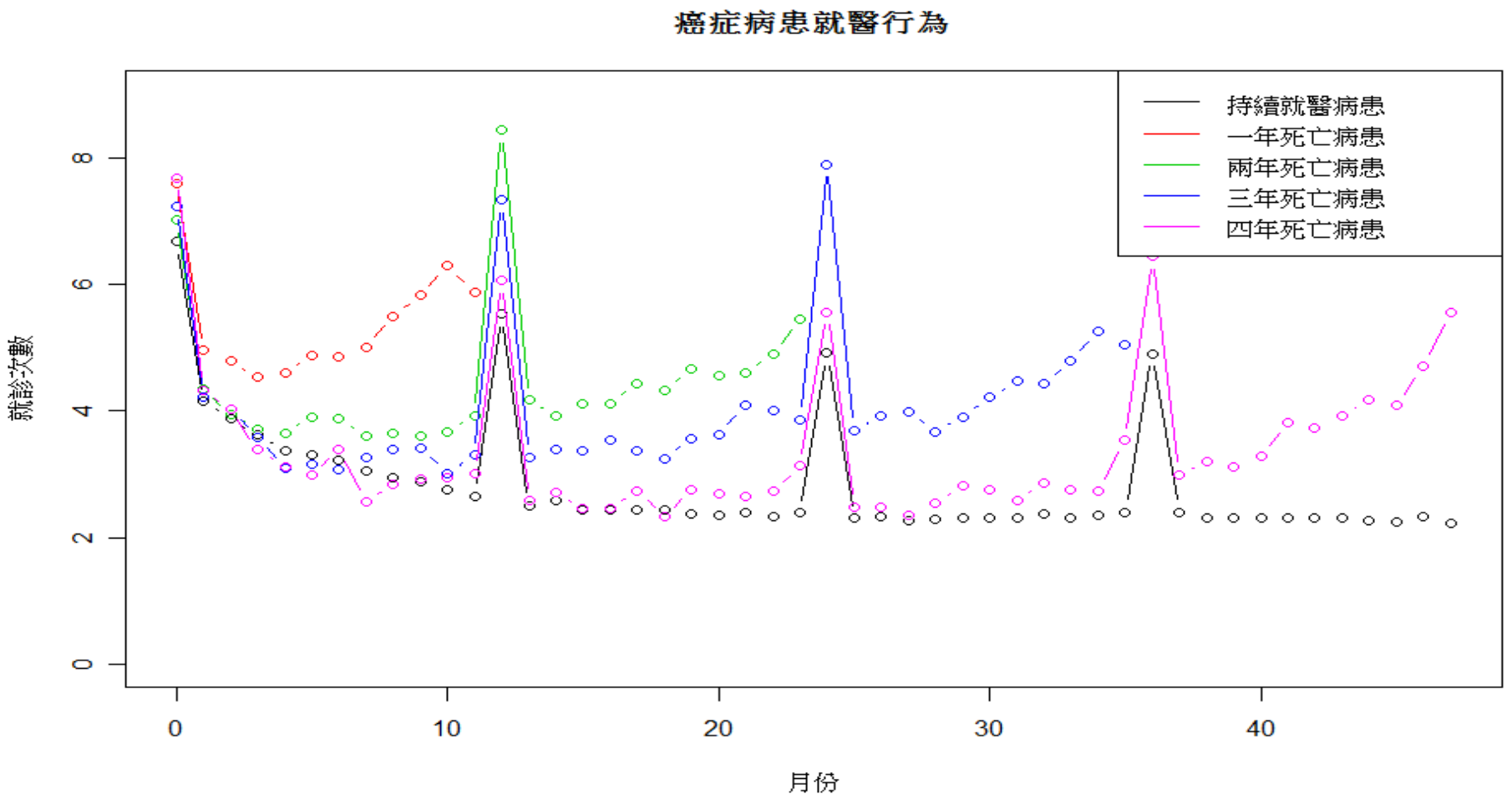
套用癌症患者的就醫特性

- 癌症屬於重大傷病的一種，需要持續接受治療或追蹤觀察，痊癒這個名詞未必適合癌症患者，較為合適的說法為獲得控制。

→因此若癌症患者連續幾年都未就醫，特別在積極就醫後突然失蹤，死亡的可能性很高。

- 另外，我國習俗多半不希望家人死於醫院，許多人在確定醫療無效後多半選擇出院回家，在健保記錄上標示為「退保」的比例頗高。

創意思考-連續兩年未就醫者認定死亡



Is this the place to learn about mining?

"THAT'S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"

案例二、921震災與中老年人死亡風險

921大地震

- 臺灣位處環太平洋地震帶，西部因人口稠密，若震央在西部可能使得災情更加嚴重。
- 921大地震發生在1999年9月21日凌晨，是臺灣近五十年來最大的地震。芮氏規模7.3級，對南投縣與台中縣造成非常嚴重的衝擊。

	死亡（含失蹤）		重傷		房屋全倒		房屋半倒	
	人數	(%)	人數	(%)	戶數	(%)	戶數	(%)
南投縣	886	36.69%	678	47.05%	23,127	52.16%	16,792	40.33%
台中縣	1,154	47.78%	411	28.52%	16,861	38.02%	12,341	29.64%
彰化縣	28	1.16%	11	0.76%	1,048	2.36%	3,054	7.34%
其它縣市	347	14.37%	341	23.66%	3,302	7.45%	9,446	22.69%
全台合計	2,415	100%	1,441	100%	44,338	100%	41,633	100%


120.78
23.85
7.3
8.0


'99 10 30

相關文獻回顧

- 相關文獻多集中在兩大方向：

一、震災後看診疾病型態與醫療利用變化

二、災難發生後短期內死亡率估算。

→越接近震央、老人與孩童死亡率越高(Chan et al., 2003; Liang et al., 2001; Liao et al., 2003)

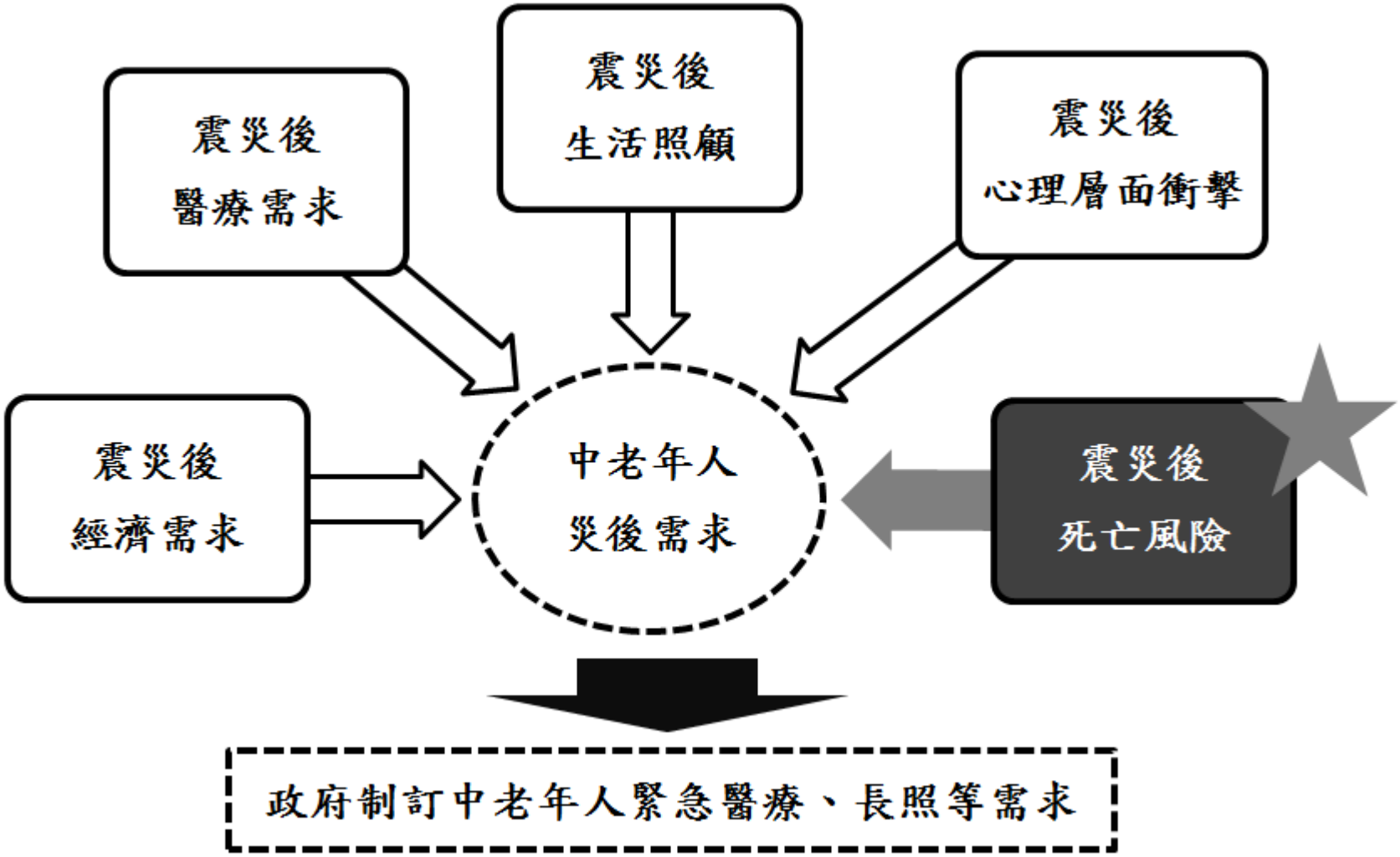
→災前已罹患精神病、住院或所得較低、災後重傷或外傷的民眾，災後一個月死亡率較高(Chao et al., 2004）

→學者研究南投縣與台中縣主要災區，發現災民於災後6個月死亡率有明顯降低，可惜缺乏對照組，且未能辨別災民身分，未能凸顯實際受創災民的情形(Chan et al., 2003)

相關文獻回顧

- 中老年人是高風險群，生理等機能退化，易受天災與氣候變化影響而患病，致使醫療與復健需求高於一般民眾（Armenian et al., 1998；Liang et al., 2001）。
- 老年人的反應比較遲鈍，可能會因災難發生後不願離開災區，或因搬到新生活環境適應不良，將造成老人的醫療照護需求高於其他年齡層(Friedsam, 1960)：

定義問題


震災後
醫療需求
震災後
生活照顧
震災後
心理層面衝擊
震災後
經濟需求
中老年人
災後需求
震災後
死亡風險
政府制訂中老年人緊急醫療、長照等需求

資料背景介紹

- 利用1998~2004年全民健保資料庫，研究對象鎖定50歲以上民眾，居住在受災最嚴重台中縣與南投縣(實驗組)，另以鄰近受災較輕微的彰化縣為對照組
- 研究目標：比較實驗組、對照組死亡風險

→年齡別死亡率、平均餘命，檢視921地震對災民死亡風險影響的持續時間。

如何確定民眾住在災區？

- **災民認定以領取921震災卡為依據**。
- 政府於災後一個月對受災戶提出健保就醫優惠方案，只要被保險人或其依附眷屬中，有任一人因震災死亡、重傷，或是住所全倒、半倒者，即可領有震災卡，並享有醫療費用、部分負擔、住院膳食之優惠。
- 根據政府統計，921災區共發出35.4萬張震災卡。
- **健保資料庫**的申報類別、給付類別、部分負擔代號等欄位，**可辨別出持震災卡**就醫的災民

如何確定921地震時常住中彰投人數？

- 戶籍人口與常住人口的差異

→1991年戒嚴令解除、經濟起飛使得差異擴大；

→1997年廢除年終戶口校正，差異問題更嚴重；

→2010年筆者親訪八八風災災區，發現差異超過50%。

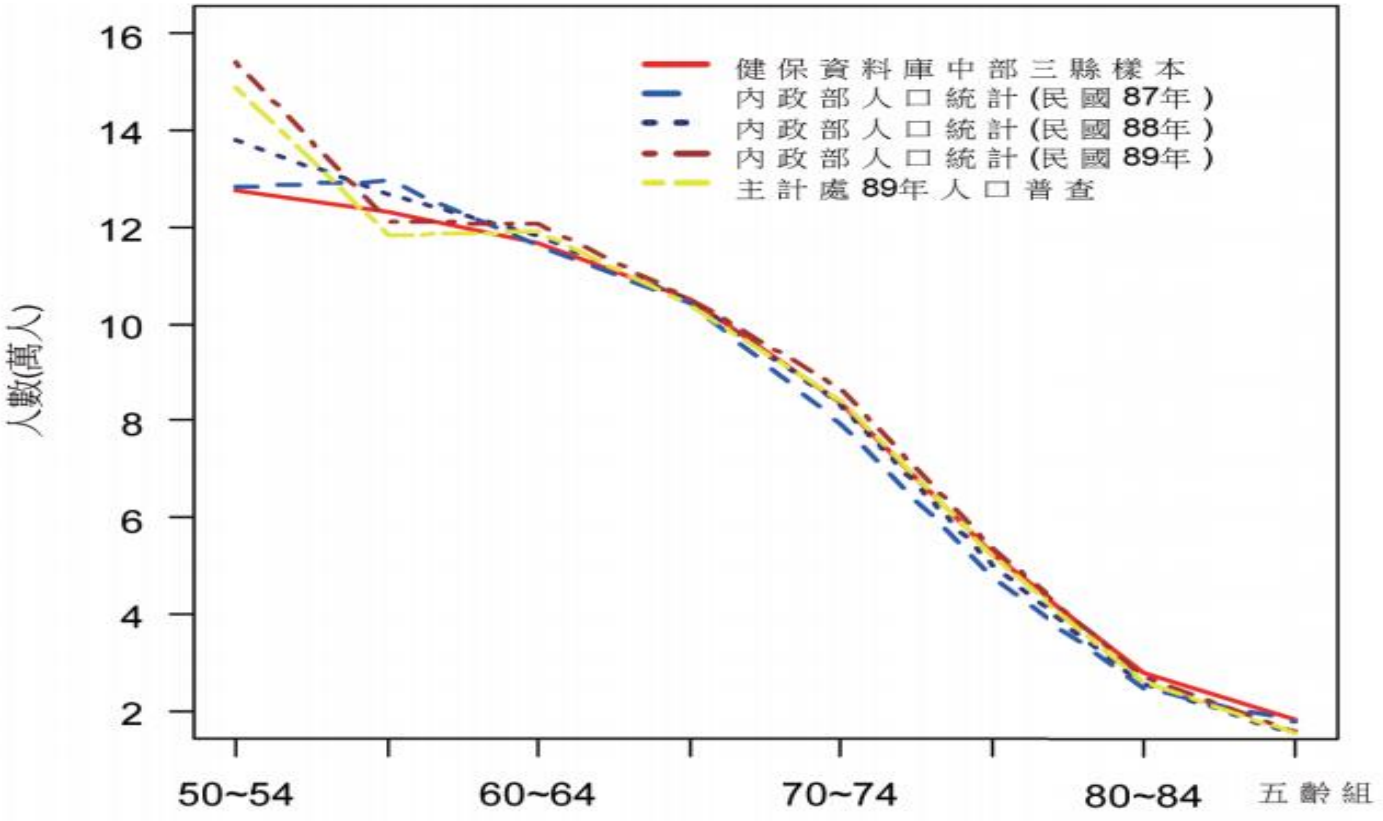
- **創意思考！以健保資料庫的輕微疾病就醫地當作常住地**

→林民浩等人（2011）以看感冒就醫地推估七成以上被保險人的常住地；

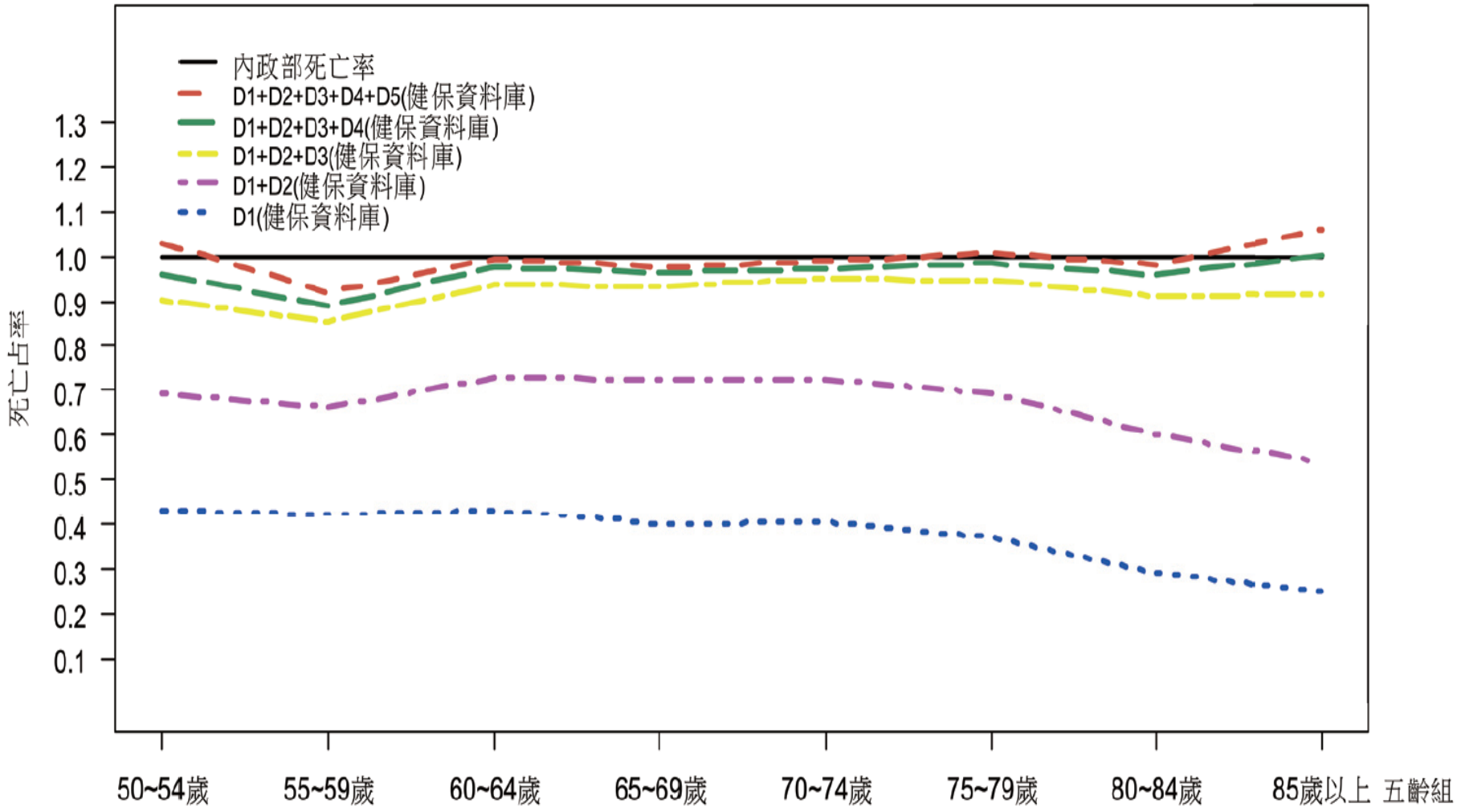
→筆者曾類似分析，發現近九成國人在三年內至少會因為感冒就醫一次，而且不到一成的感冒患者會到不同縣市就醫。

常住人口估算結果

- 以健保資料庫推算50歲以上居住在中彰投人口達65.5萬人，與戶籍登記、戶口普查相比，誤差在2%以內。



單純用退保別判斷死亡？


內政部死亡率
D1+D2+D3+D4+D5(健保資料庫)
D1+D2+D3+D4(健保資料庫)
D1+D2+D3(健保資料庫)
D1+D2(健保資料庫)
D1(健保資料庫)
死亡占率
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
50~54歲
55~59歲
60~64歲
65~69歲
70~74歲
75~79歲
80~84歲
85歲以上
五齡組

估算死亡率

- 粗死亡率=死亡人數/總人數。
- 調整後死亡率(Direct Method Adjustment；ADR)

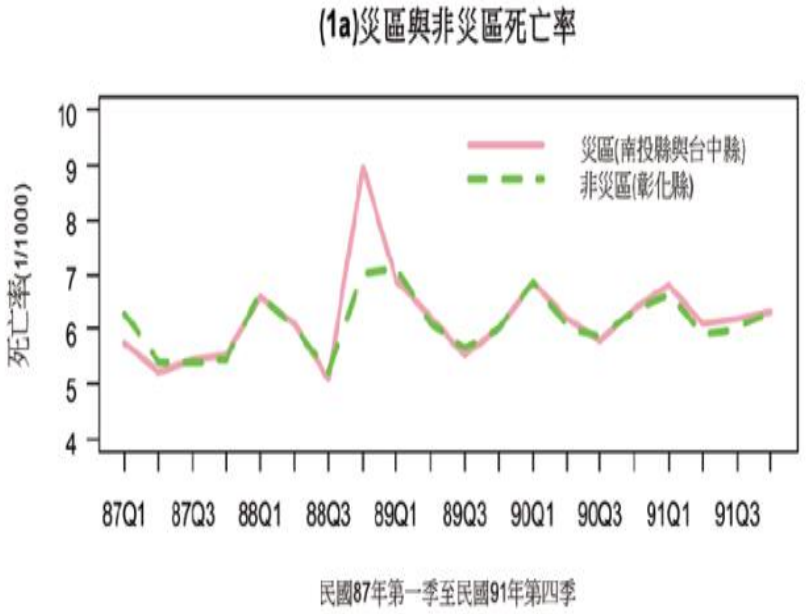
→以全國人數為標準母體，按照人口結構調整死亡率。

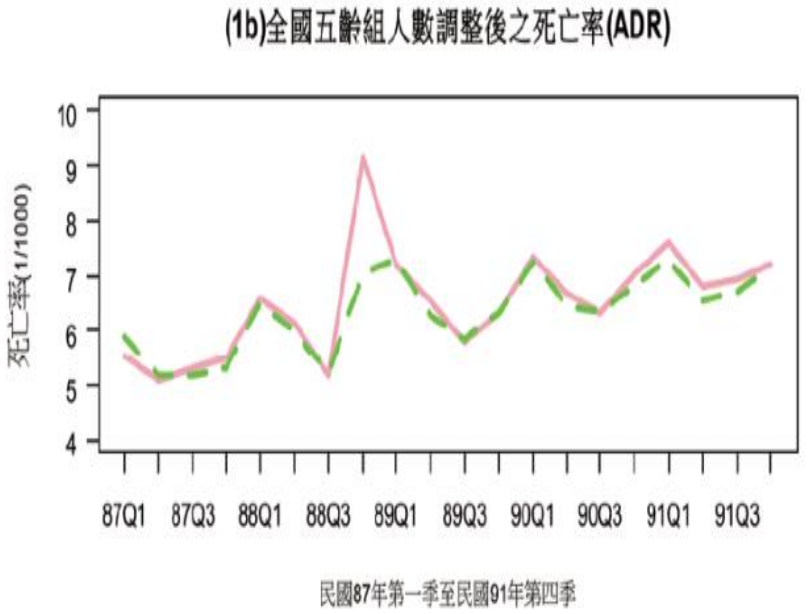
- 標準化死亡比(Standardized Mortality Ratio；SMR)

→ 各地區標準死亡率/母體標準死亡率。

由災區別看死亡率

- 每年度Q1死亡率高。時值冬天氣候寒冷，老人比較容易因感冒引起肺炎或其他疾病導致死亡。
- 震災發生前(87Q1～88Q3)，災區與非災區的死亡率幾乎沒有差異。
- 發生當季(88Q4)，明顯看出災區死亡率跳升。
- 發生後第一季(89Q1)災區死亡率回復如非災區水準。

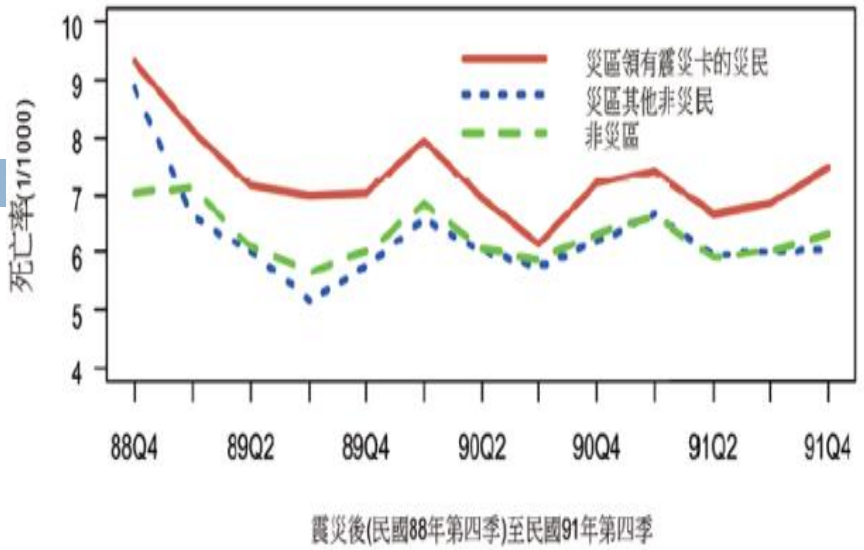




震災卡別看死亡率

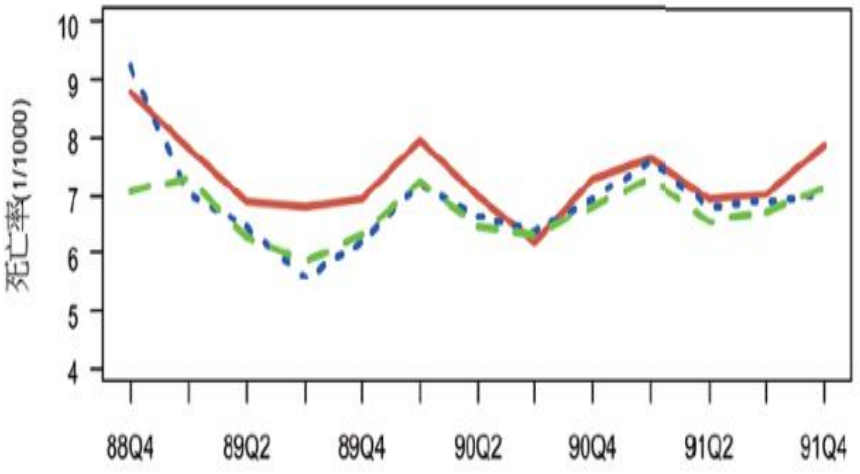
- 震災對災民的死亡率影響至少兩年。
- 災後第1年(89Q1～90Q1)直到次年春節過後，災民死亡率明顯較高，直到90Q2才開始減緩，但仍比非災區、或災區非災民高。

(2a)災區領有震災卡的災民、災區其他非災民與非災區死亡率



震災後(民國88年第四季)至民國91年第四季

(2b)全國五齡組人數調整後之死亡率(ADR)



震災後(民國88年第四季)至民國91年第四季

案例三、臺灣女性戶長

臺灣女性戶長比例攀升

- 台灣家戶中女性戶長的比例在十年內上升了10%，在不同類型的家戶中成長率不同。

→哪些因素與女性擔任戶長有關？

1990與2000年戶口普查台灣女性戶長比例

	總計	核心家庭	三代家庭
1990女性戶長比例	23.4%	19.6%	24.8%
2000女性戶長比例	33.4%	28.8%	30.6%

資料背景介紹

- 以台灣地區家戶屬性及戶長性別為分析對象，希冀藉此瞭解家庭中戶長的性別而言，是否存在男女不平等的現象。
- 2000年台灣地區戶口普查家戶資料，依地區：

→北部(台北市、桃園縣)：1,008,718戶。

→中部(台中縣、市)：551,191戶。

→南部(高雄縣、市)：644,671戶。

→東部(花蓮縣、台東縣)：118,294戶。

變數定義及說明

- 反應變數：

  ── 戶長性別

- 解釋變數：

  →戶長年齡、戶長婚姻狀況、戶長教育程度、戶長是否為原住民、家計負責人性別、家戶型態。

解釋變數類型

	性別	婚姻狀況	教育程度	是否為原住民	家戶型態
1	男	未婚	自修，國小	是	夫婦兩人
2	女	有配偶或同居	國中，高中職	否	父母及未婚子女
3		已離婚或分居	大專		父(或母)及未婚子女
4		配偶死亡	碩博士		祖父母，父母及未婚子女
5			不識字		父母及已婚子女
6					祖父母及未婚子女
7					單身家戶
8					其他家戶：有親屬關係
9					其他家戶：無親屬關係

註：「戶長年齡」為連續變數。

使用的分析模型

- 本問題屬於監督學習(Supervised Learning)，已知反應變數的數值，尋求與解釋變數間的關係。
- 常用的分類方法包括：（可使用R軟體）
  - 羅吉斯迴歸(Logistic Regression)
  - 分類與迴歸樹(Classification and Regression Tree)
  - 類神經網路(Neural Network)
  - 支持向量機(Support Vector Machine)

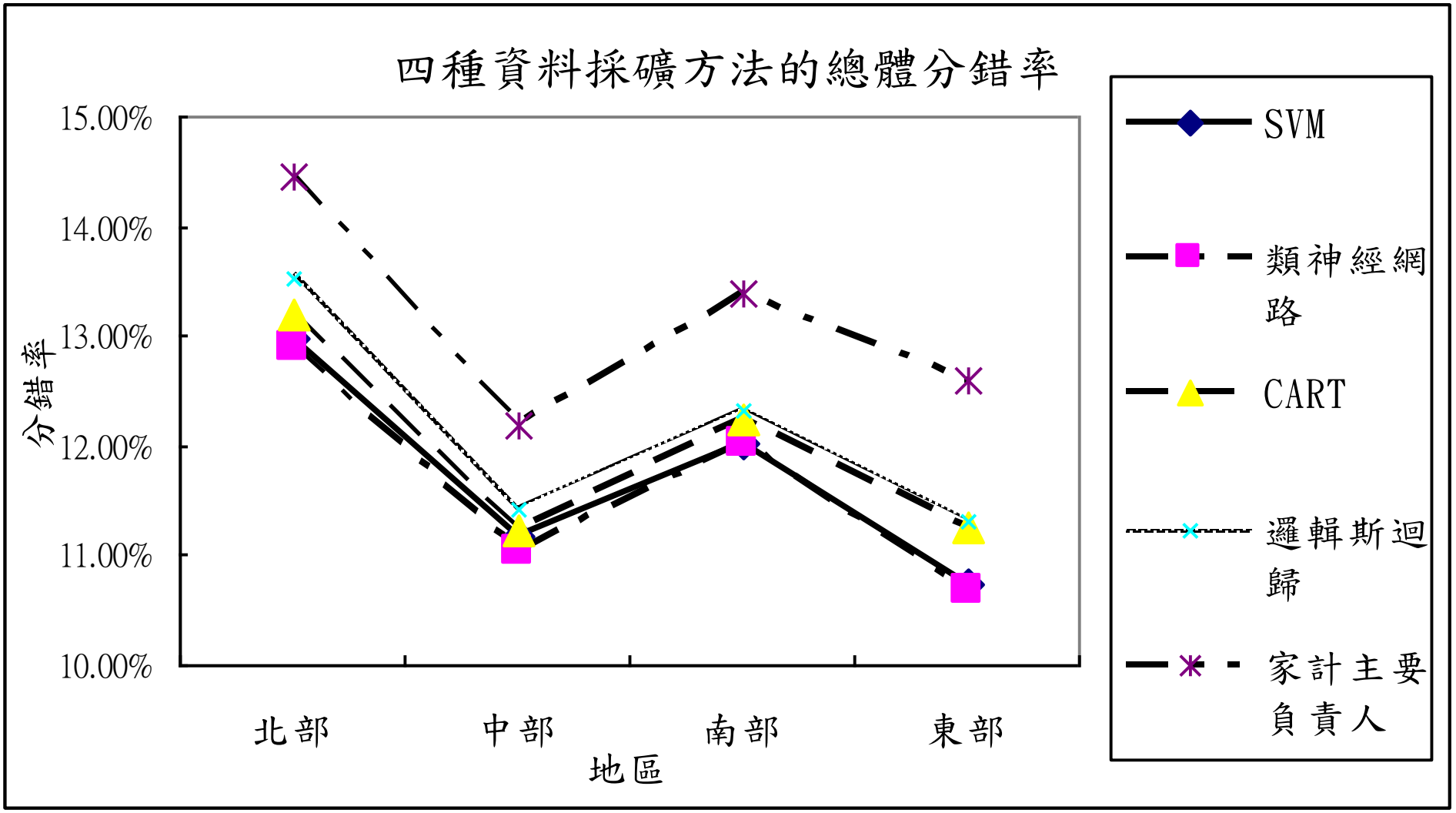
判斷標準

- 以戶長性別分類錯誤的比例為判斷標準：

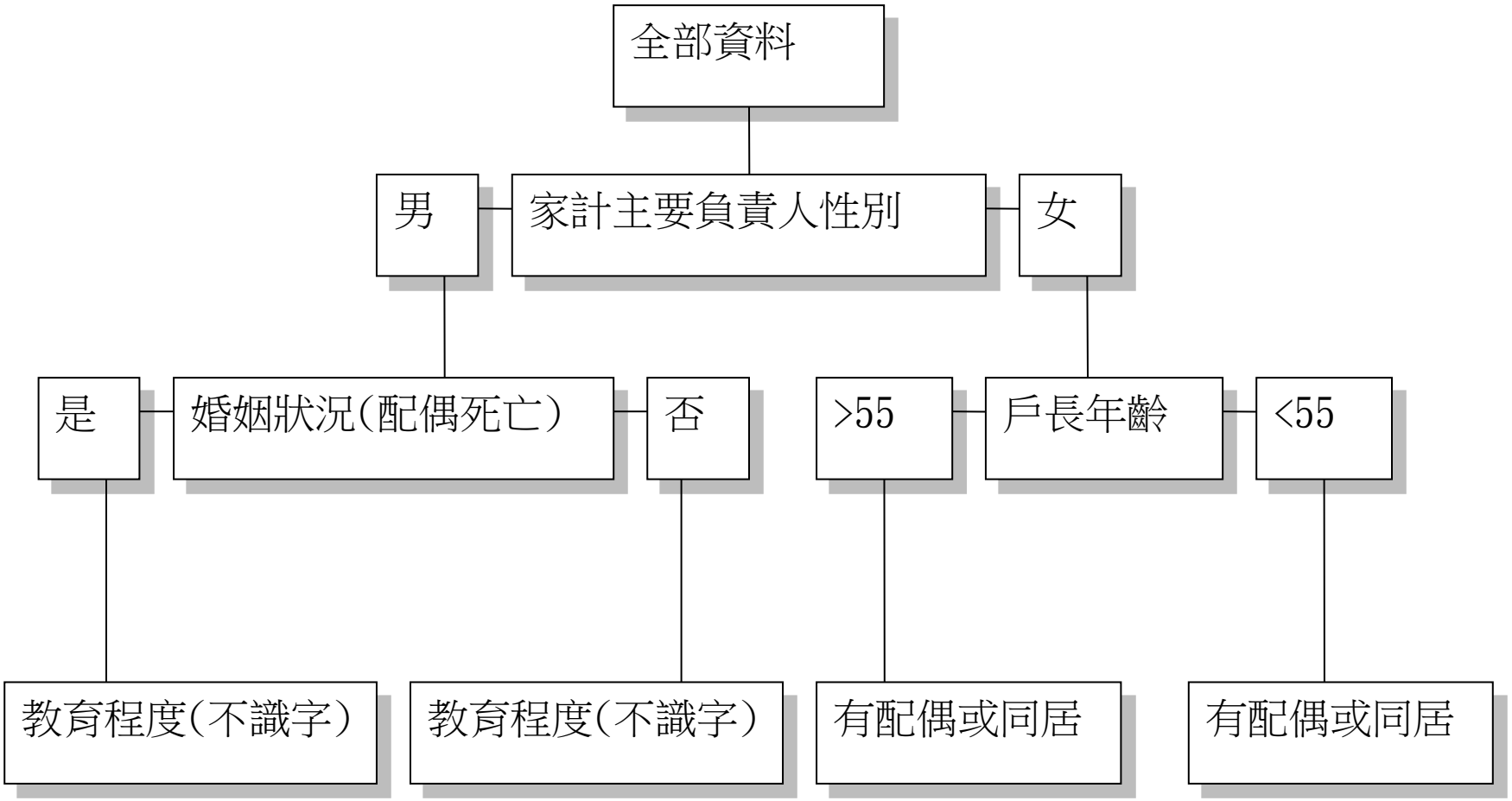
		實際值	
		女	男
預測值	女	A	B
	男	C	D

$$\text{分錯率} = \frac{B+C}{A+B+C+D}$$

四種模型的比較


四種資料採礦方法的總體分錯率
15.00%
14.00%
13.00%
12.00%
11.00%
10.00%
分錯率
北部
中部
南部
東部
地區
SVM
類神經網路
CART
邏輯斯迴歸
家計主要負責人

CART的分類結果


全部資料
男
家計主要負責人性別
女
是
婚姻狀況(配偶死亡)
否
>55
戶長年齡
<55
教育程度(不識字)
教育程度(不識字)
有配偶或同居
有配偶或同居

羅吉斯迴歸的顯著變數

羅吉斯迴歸重要變數	北部	中部	南部	東部
家計主要負責人性別	40.241	49.528	39.099	47.269
未婚	16.598	20.915	18.598	29.973
有配偶或同居	13.934	18.498	14.152	17.093
碩博士	11.434	13.674	14.795	12.382

註：表中為四地區指數轉換轉換後的估計值。

使用部分資料的電腦模擬

- 當處理的資料量龐大，在不損失分類精確度的考量下，減少資料分析的運算量：

  SVM →6小時 (僅使用10%資料)

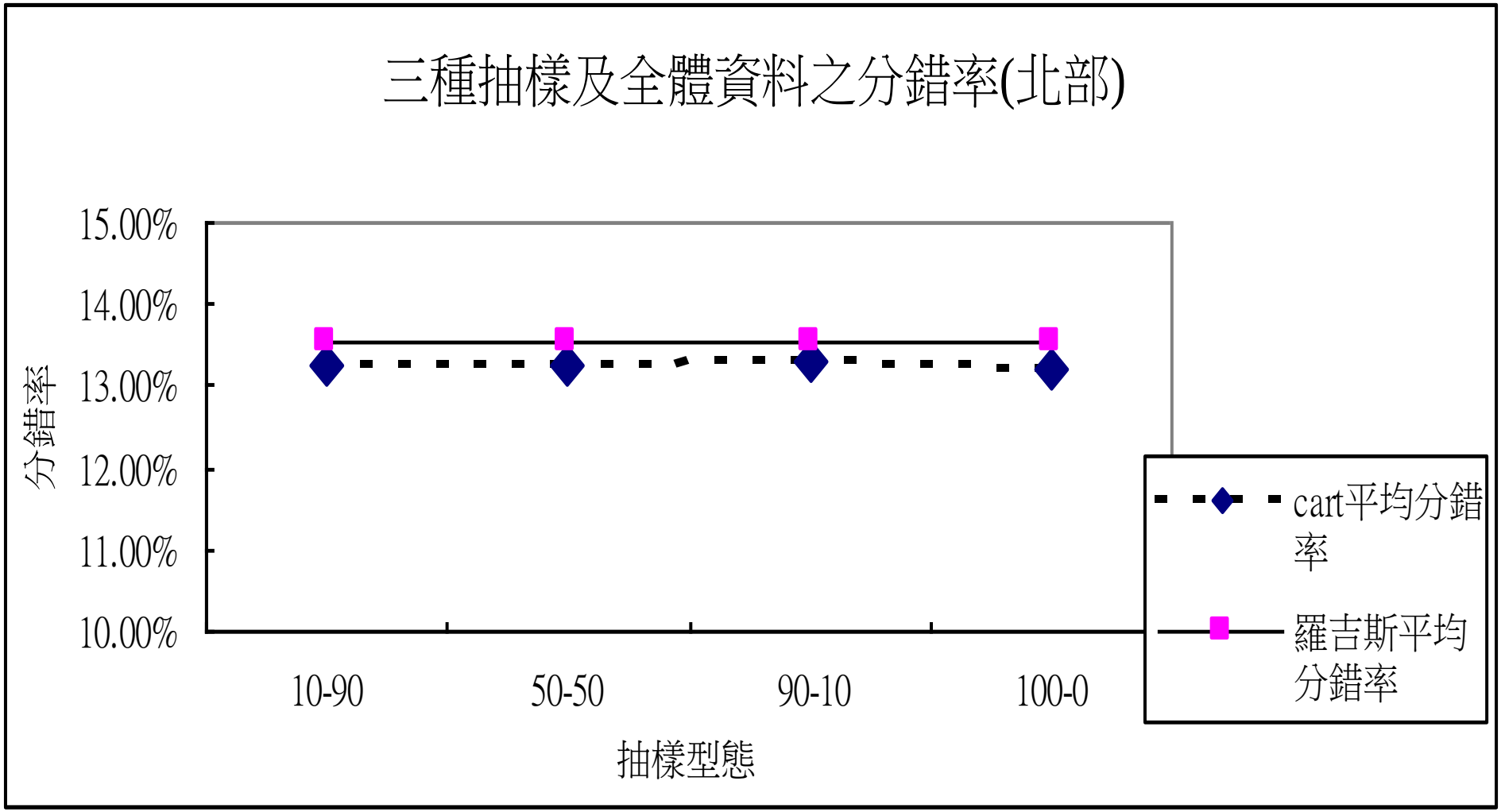
  類神經網路→2小時

  CART →3分鐘

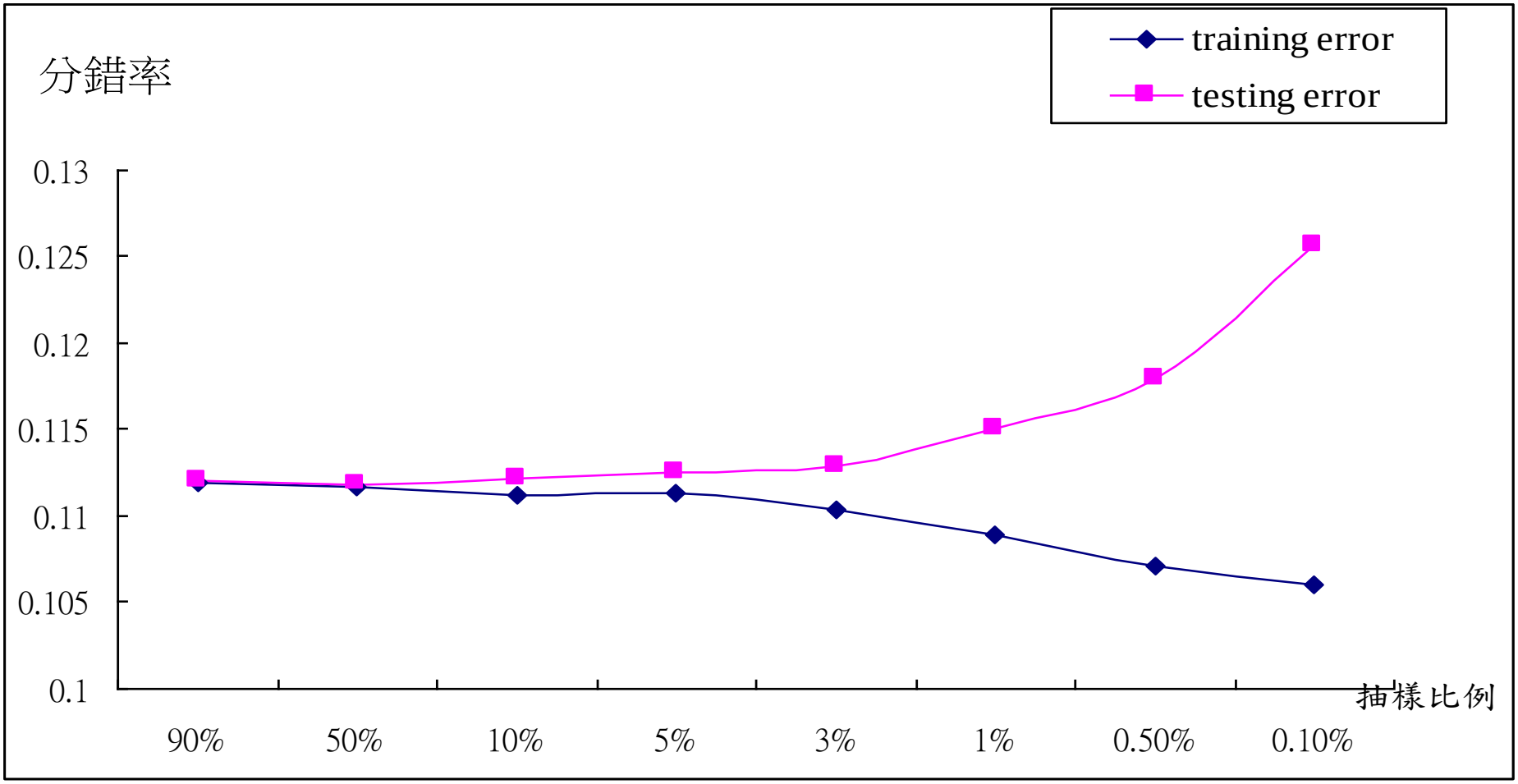
  羅吉斯迴歸→1分鐘

- 隨機將四地區固定比例的資料建立模型，比例分別為10%、50%、90%，重覆模擬20次，以CART和羅吉斯迴歸為例（需時較短）。

交叉分析的分類結果


三種抽樣及全體資料之分錯率(北部)
15.00%
14.00%
13.00%
12.00%
11.00%
10.00%
分錯率
10-90
50-50
90-10
100-0
抽樣型態
cart平均分錯率
羅吉斯平均分錯率

完全隨機抽樣的結果


分錯率
training error
testing error
0.13
0.125
0.12
0.115
0.11
0.105
0.1
90%
50%
10%
5%
3%
1%
0.50%
0.10%
抽樣比例